Zero-shot Detection of Out-of-Context Objects Using Foundation Models

Anirban Roy [1], Adam Cobb [1], Ramneet Kaur [1], Sumit Jha [2], Nathaniel D. Bastian [3],
Alexander Berenbeim [3], Robert Thomson [3], Iain Cruickshank [3],
Alvaro Velasquez [4], Susmit Jha [1]
[1] SRI, [2] Florida International University, [3] United States Military Academy, [4] DARPA

Abstract

We address the problem of detecting out-of-context (OOC) objects in a scene. Given an image, we aim to detect whether the image has objects that are not present in their usual context and localize such OOC objects. Existing approaches for OOC detection rely on defining the common context in terms of the manually constructed features, such as the co-occurrence of objects, spatial relations between objects, and shape and size of the objects, and then learning such context for a given dataset. But context is often nuanced ranging from very common to very surprising. Further, learned context from specific datasets may not be generalized as datasets may not truly represent the human notion of what is in context. Motivated by the success of large language models and more generally, foundation models (FMs) in common sense reasoning, we investigate the FM's ability to capture a more generalized notion of context. We find that a pre-trained FM, such as GPT-4, provides a more nuanced notion of OOC and enables zero-shot OOC detection when coupled with other pre-trained FMs for caption generation such as BLIP-2, and image in-painting with Stable Diffusion 2.0. Our approach does not need any dataset-specific training. We demonstrate the efficacy of our approach on two OOC object detection datasets, achieving 90.8% zero-shot accuracy on the MIT-OOC dataset and 87.26% on the IJCAI22-COCO-OOC dataset.

1. Introduction

Integrating context in machine learning models is crucial to fostering robust inference capabilities, particularly in tasks like object detection in a scene. Typically, objects in natural images appear in an appropriate context. These contextual cues can aid object detection [4, 7, 23, 33, 42, 52]. However, incorrect contextual cues can adversely affect the performance of object detection in both humans [7, 38, 49] and machine learning models [30, 38].

Thus, creating robust and reliable object detection systems requires identifying and appropriately handling objects appearing in atypical contexts - scenarios where predictions may not exhibit expected reliability. While detecting in-context objects is extensively studied [3, 5, 17, 28, 44], the problem of detecting out-of-context (OOC) objects [1, 11] is not thoroughly explored and existing approaches have relied on dataset-specific training. In this work, we use FMs for zero-shot OOC detection without any retraining or fine-tuning to learn the usual context relations.

Figure 1. Left: An OOC sample from MIT-OOC dataset. Right: Generated non-OOC image by removing the OOC object by in-painting the image.

The quintessential example of an out-of-context object is the proverbial *elephant in the room* [38]. In Figure 1 (left), we show a similar example from the MIT-OOC dataset first studied in [11]. It would be extremely out of context for an elephant to be perched in a nest atop a tree. When a caption generation model, BLIP-2 [26], is queried with this image, it generates the caption "a large elephant statue sitting in a tree with a nest". An elephant statue nestled in a tree is also unusual but perhaps less contextually jarring than an actual elephant. When GPT-4 [34] is queried with this caption, it identifies the caption as unusual and indicative of an out-of-context scene. By applying an image in-painting model, Stable Diffusion 2 [37], to mask and complete parts of the image, we successfully create an image

Figure 2. OOC images from the MIT-OOC dataset [11]. The top two images are out of context with a couch on the top of a floater in a swimming pool, and an elephant in the room. The bottom row images are more subtle and not as surprising as the samples in the top row. For example, airplanes flying at a low altitude over a beach are usually rare but common at Maho Beach.

Figure 1 (right) where the OOC object has been eliminated, and the caption generated by BLIP-2 is deemed normal by GPT-4. This enables localization of the out-of-context object(s) in the image.

While supervised learning of context for out-of-context detection [1] has traditionally framed the contextual appropriateness of an object as a binary concept, the context of an object is more nuanced and subjective. We illustrate this in Figure 2 using a set of samples from the MIT-OOC dataset. The notion of an object's 'usual context' is not a dichotomous concept, and even humans can sometimes grapple with accurately discerning context [7]. The detection of out-of-context objects necessitates a comprehensive understanding of what is common or contextually consistent in a scene. Existing approaches strive to learn this common context from a finite labeled image dataset, where context is defined by a predetermined set of relations, such as object co-occurrence, spatial relations between objects, and physical properties like shape and size. However, comprehensively enumerating all common contextual relations and learning these from a limited image dataset is daunting. We identify the following primary challenges in out-of-context object detection:

- Context is a complex, subjective, and multifaceted concept that is not simply binary. The detection of out-of-context objects necessitates a comprehensive understanding of the typical context for a variety of objects.
- Learning context from a small dataset faces the risk of missing different possibilities and variations in the context that may not be represented in the dataset.
- Hand-engineered features such as co-occurrences, spatial relations, and relative sizes may not fully capture what is in context and what is not.

Contributions. We posit that the detection of out-of-context (OOC) objects can be framed as a multimodal problem, requiring a language-guided understanding of the context for accurate image analysis. This hypothesis serves as the motivation behind our utilization of foundational models for OOC detection. We use a large language model as the source of contextual knowledge to detect OOC objects in images. Our OOC detection pipeline is presented in Figure 3. We identify the key novel contributions below:

- We utilize the GPT-4 to capture a comprehensive understanding of context. Instead of learning context from a limited, human-curated dataset, we rely on FMs trained on a massive quantity of multimodal data to establish what is commonplace and normal context.
- By employing FMs and vision-language models, we encapsulate a broad understanding of context, transcending the restrictions of hand-engineered features like co-occurrences, relative sizes, and spatial relations. We demonstrate that the pre-trained LLM's innate understanding and common sense knowledge are effective in capturing the notion of the usual context of objects.
- We propose a novel pipeline comprising pre-trained multimodal models for scene understanding and caption generation, natural language understanding to discern whether a caption describes an out-of-context scene, and image in-painting for counterfactual image generation. This pipeline is capable of zero-shot OOC detection and localization without any dataset-specific training.
- We evaluate our approach on two public datasets for OOC detection: MIT-OOC dataset [11] and IJCAI22-COCO-OOC dataset [1]. In comparison to the SOTA method's reliance on dataset-specific training, our zero-shot approach using FMs achieves a high OOC detection accuracy without any dataset-specific training (90.9% vs 73.3% on MIT-OOC dataset and 87.26% vs 84.85% on IJCAI22-COCO-OOC dataset where SOTA methods use a manual definition of context and learning on the dataset). When compared to the baseline of visual question-answering where we directly query BLIP-2, our accuracy is 90.8% (ours) vs 23.4% (BLIP-2) on the MIT-OOC dataset and 87.26%(ours) vs 26.8% (BLIP-2) on the IJCAI22-COCO-OOC dataset.
- Not only does our approach accurately detect OOC objects, but it also quantifies context and provides a human-interpretable natural language explanation for the result using FMs.

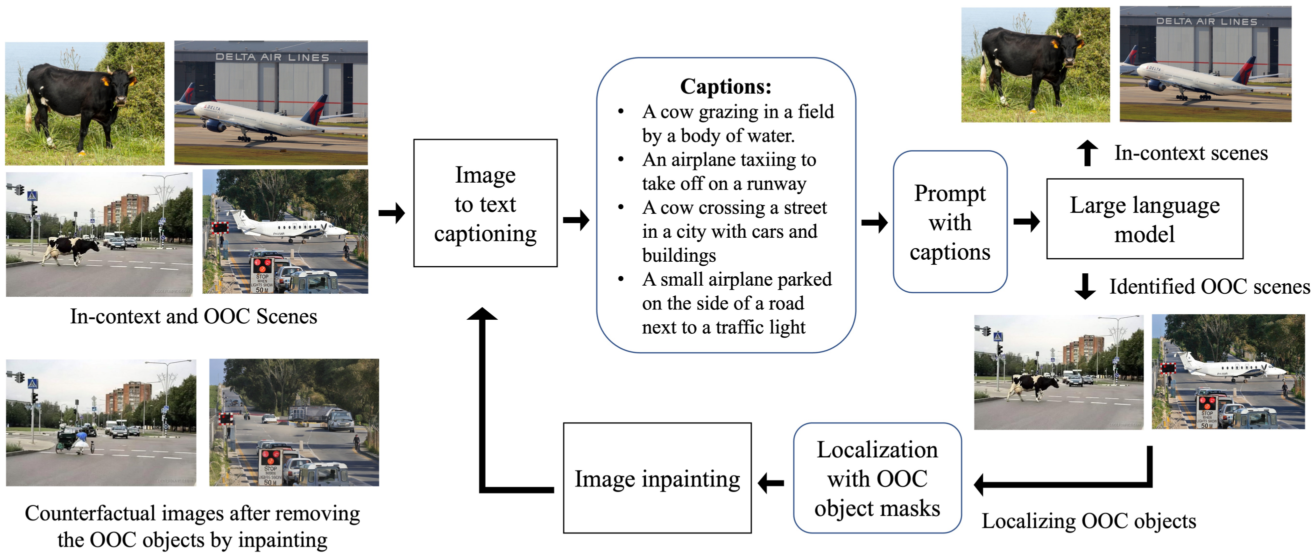

Figure 3. The overview of the proposed zero-shot OOC detection method using FMs. The image captions generated from BLIP-2 are used to prompt GPT-4 on whether the description is a normal context. For the identified OOC scenes, a candidate OOC object is determined. A segmentation mask corresponding to the object guides the diffusion-based in-painting to generate counterfactual images. The counterfactual images are further examined to detect the OOC scenario. For example, the OOC cow gets in-painted with a vehicle and the OOC airplane gets in-painted with road and other vehicles. As in-painting objects make the updated image in context, we localize these objects as OOC. The OOC images are shown on the bottom right and the in-painted images are shown on the bottom left.

2. Related Work

Context for object detection. Contextual cues are important for detecting objects in images. Common contextual cues include co-occurrences of objects, object attributes, and spatial relations between objects [23, 52]. While accurate contextual cues are shown to be useful for object detection [4, 7, 23, 33, 42, 52], inaccurate contextual cues can negatively impact the performance of object detection [7, 30, 38, 49]. Thus, detecting OOC scenarios is important to develop a reliable object detection model.

Out-of-context object detection. [11] introduced OOC detection as a scene understanding task. They capture contextual information using global scene categorization, object co-occurrences, and geometric context such as relative positions and scales of objects. [7] consider neural network-based models to learn contextual relations in a data-driven manner. Some approaches consider context as the generic background and do not exploit informative cues such as label dependencies and relative object properties in the image [4,49]. [1] considered a graph contextual reasoning network (GCRN) to detect OOC objects. GCRN consists of two separate graphs: a representation graph to learn object features based on the neighboring objects, and a context graph to explicitly capture contextual cues from the neighboring objects. These approaches rely on learning context from a relatively small dataset. We use pre-trained foundation models for OOC detection and demonstrate comparable or better performance to the SOTA methods without explicit learning of context relations.

Vision-language models. Vision-language models aim to learn a representation to align visual information and natural language to perform various downstream tasks such as image captioning [20, 25, 26], visual question answering [2, 14], and grounded caption generation [15]. As the size of the dataset and the models are increasing, recent approaches focus on learning representation with frozen vision and/or language backbones. It is usually common to consider a frozen vision model and adopt a language model [9, 27, 35, 50] as visual features from a well-trained vision model can provide informative visual cues. However, due to the growth of large language models [12, 34, 45, 51], learning with frozen language models too is becoming popular where a transformation layer is learned to bridge the gap between frozen visual and language features [8, 14, 19, 20, 31, 43]. Despite the success of vision-language models, it is not yet feasible to directly query such a model with a prompt to determine whether an image is out of context. We experimentally demonstrate this in our evaluation. This motivates our approach connecting the vision-language model BLIP-2 with GPT-4.

Commonsense in foundation models and their use for zero-shot tasks. FMs such as GPT-3 and GPT-4 [34, 36] currently provide the SOTA performance in tasks such as natural language translation, predicting long-range text dependencies and even translation to structured representations such as program synthesis. These models are capable of simple reasoning tasks [22, 41] with some reservations [32, 46]. FMs can be made to generate a coherent chain of thought - a series of short sentences that mimic the reasoning process of a person when responding to a question. We exploit this common sense reasoning capability of FMs. [22] shows impressive zero-shot FM performances on diverse tasks including arithmetic and symbolic reasoning, and other logical reasoning tasks such as tracking shuffled objects. FMs have also been used for zero-shot retrieval [2, 40] and zero-shot visual question-answering [19, 26, 43], zero-shot image captioning [2, 26]. [18] used FMs for detecting five types of traffic-related anomalies with carefully designed prompts. Our work is the first to demonstrate the use of FMs for zero-shot OOC detection in general scenes.

3. Technical Approach

The task of detecting out-of-context (OOC) objects demands a comprehensive understanding of what is considered commonplace or contextually consistent. We leverage the knowledge embedded within foundational models (FMs) that have been exposed to a vast array of multimodal data from varied sources. This allows us to detect OOC objects in a zero-shot fashion without any retraining or finetuning. For an input image, we aim to detect whether the objects in the image are in context or not. If the image is deemed out of context, we proceed to localize the OOC object(s). We employ an FM to provide a quantitative OOC score along with an explanation derived from the FM.

Given an image, we first generate a caption using BLIP-2 [26], a multimodal FM, detailing the objects and their relationships in the image. We query GPT-4 [34] to ascertain whether the object configurations in the image, as depicted by the caption, are commonplace or if any objects appear out of context. To localize the OOC object, we generate counterfactual images by removing the possible OOC object(s) from the image by in-painting the object(s) using Stable Diffusion 2 [37]. If the omission of the object(s) from an image renders the caption describing the image appear in context to GPT-4, we conclude the object(s) is indeed an OOC object. Given that the foundational models are trained on an extensive range of multimodal data, they encapsulate the concept of context and enable zero-shot detection of out-of-context images. These models can be easily adapted to question-answering for OOC detection without requiring finetuning. Consequently, our approach does not require training on any specific dataset for OOC detection. The overall process is illustrated in Figure 3. We describe the primary components below.

3.1. Image captioning.

The first step of our approach uses an FM BLIP-2 [26] to get an informative caption to describe an image and depict the relations between the objects. BLIP-2 consists of a frozen vision encoder and a frozen language model. A transformer [47] network is learned to align visual features and text features to perform downstream task tasks such as image captioning and visual question answering. Owing to its training on a large-scale dataset of 234 million images that are curated from various datasets such as COCO [28], Visual Genome [24], and LAION [39], BLIP-2 achieves state-of-the-art performance on zero-shot image captioning. In the majority of the cases, these captions capture the key details needed to detect OOC images and we demonstrate this in our experimental evaluation. Furthermore, we analyze cases when the captions miss OOC objects or fail to capture OOC attributes of an object.

3.2. Prompting the foundational model.

Given the generated caption, we prompt an FM to check whether the caption is uncommon and the image is OOC. We use GPT-4 in our approach. GPT-4 is a large language model that is trained with a vast amount of multimodal data and is good at following instructions. The model has achieved impressive performance in various tasks including science and art exams, interactive question answering, and reasoning [34]. The model is known to have captured extensive common knowledge in various domains. Thus, we leverage this model to detect uncommon or abnormal scenarios as described in the caption[1] to detect OOC images. Specifically, given the captions, we ask two queries to the model with the following prompt templates that were selected via manual exploration:

- Prompt1: `<caption> - How normal is this on a scale between 0 to 10?`
- Prompt2: `<caption> - How common is this on a scale between 0 to 10?`

Interestingly, GPT-4 not only provides a binary (yes or no) response but often qualifies its response with conditions that could render the scenario normal or common. The model considers plausibility when asked about normality and frequency when asked about commonality. For example, flood is plausible but it does not occur frequently. We explored other prompt options including ‘How likely’, ‘How uncommon’, and ‘Is this out-of-context’. We find that Prompt2 is most effective for OOC detection.

3.3. Localizing OOC object in images.

Once the image is OOC, the next step is to localize the part of the image that has OOC object(s). We perform coun-

[1]We only have access to the language input API to GPT-4 but we use BLIP-2 as a baseline for a direct visual Q&A approach.

terfactual evaluations using the Stable Diffusion 2 [37] as the in-painting model. This model, which strikes a balance between reducing image generation time and preserving details, generates high-fidelity images.

The pipeline for OOC object localization is shown in the bottom part of Figure 3 when an image is deemed as OOC. We first determine the OOC object candidates from the GPT response. Then we use a prompt-based segmentation model [21, 29] to generate the masks corresponding to the OOC object. Then, we generate the counterfactual image by infilling the object using Stable Diffusion 2 [37]. Finally, the infilled image is again fed to the multimodal FM to generate the caption. The caption is examined by the GPT-4. If the counterfactual image is determined as in-content, then the object is marked as an OOC instance.

4. Experiments

In this section, we describe the dataset, explain the experimental setup, present results, perform error analysis, and discuss limitations. We evaluate in terms of detecting OOC images and localizing the OOC object in the image.

Datasets. We consider two OOC datasets to evaluate our approach: 1) MIT-OOC [11] and 2) IJCAI22-COCO-OOC [1]. We choose these two datasets as these come with object-level annotations allowing the evaluation of localizing OOC objects: **MIT-OOC [11]** consists of human-annotated images where each image has one or more OOC OOC objects. These images present natural OOC scenarios from indoor and outdoor scenes. The objects are annotated with a polygon bounding box and an OOC tag. **IJCAI22-COCO-OOC [1]** dataset is created from COCO 2014 validation set [28] by transplanting OOC objects on images. Authors synthetically generate OOC scenarios by placing objects in images to violate common contextual relations [6, 49]. The objects are annotated with COCO-like segmentation masks. This dataset follows a similar strategy to generate OOC images as the Cut-and-Paste dataset [49]. This contains a larger number of images and considers OOC objects violating co-occurrence, location, and size relations.

Baseline: Direct visual question answering queries. In this baseline, we consider a visual question-answering (VQA) approach for determining whether an image is OOC. VQA is shown to be successful in answering questions about visual content in the image such as object attributes, spatial layout of objects, actions, and actor-object interactions. We consider BLIP-2 [26] for VQA as it achieves SOTA performance among the open-source models. We ask the following question to the VQA model: "Is this image out of context?". The results for this baseline are shown in table 1 where we compare the average accuracy (%) of detecting OOC images.

Ablation study. We investigate other visual FMs and LLMs to recognize the OOC objects. Apart from BLIP-2, we consider the InstructBLIP [13] model for captioning, and apart from GPT-4, we query Llama3 [16] to determine OOC scenarios. We consider MIT-OOC for this study as it contains diverse OOC scenarios. The results are shown in table 2. We use Prompt2 for querying. We notice that the performance of BLIP-2 and InstructBLIP are comparable and GPT-4 significantly outperforms Llama3 performance.

Dataset	VQA	Ours
MIT-OOC dataset	23.45	**90.82**
IJCAI22-COCO-OOC	26.78	**87.26**

Table 1. Comparison with the baseline.

FM combination	Accuracy
InstructBLIP + Llama3	73.85
BLIP-2 + Llama3	74.77
InstructBLIP + GPT-4	88.13
BLIP-2 + GPT-4	90.82

Table 2. Comparison with the combination of visual FM and LLMs on MIT-OOC.

4.1. Comparison with the state of the art.

We compare with the state-of-the-art (SOTA) approaches on MIT-OOC [11] and IJCAI22-COCO-OOC [1] in terms of detecting OOC images. SOTA approaches aim to capture context by defining features on objects attributes and relations between objects. Then dataset-specific statistics are learned as common context. For example, [11], defines the common context in terms of co-occurrences of objects, the geometry of objects, support relations among the objects, and global scene details. These context cues are learned from a set of in-context images from SUN dataset [48] and combined to model common context [10]. The results on MIT-OOC are shown in table 3. Our approach achieves a better performance without manually defining out-of-context and without any specific learning of a common context from a dataset as done in [11]. The results on IJCAI22-COCO-OOC are shown in table 4. [1], use a graph neural network to learn the context in terms of co-occurrence and spatial relations between objects. The common context relations are learned in a data-driven manner from in-context COCO images and the context cues are suitable for the OOC images in IJCAI22-COCO-OOC. We choose the setup where ground-truth bounding boxes are used in [1] to avoid errors due to object detection. Our approach achieves superior performance, again, without learning any dataset-specific context relations. Recall that we consider two prompts to query the GPT-4 model and find Prompt2 to be more effective. Since we prompt GPT-4 to provide a quantitative score for OOC ranging from 0 to 10, we use a threshold of 5 when detecting an OOC image.

Approach	Accuracy (%)
Combination of contexts [11] (manual definition + learning of context)	73.29
Our approach using pre-trained FMs and no finetuning	
Prompt1	80.73
Prompt2	**90.82**

Table 3. Comparison with the SOTA on MIT-OOC [11].

Approach	Accuracy (%)
Graph network [1] (learning of context)	84.85
Our approach using pre-trained FMs and no finetuning	
Prompt1	68.78
Prompt2	**87.26**

Table 4. Comparison with the SOTA on IJCAI22-COCO-OOC [1].

Localizing OOC objects in images. Besides detecting an image as OOC, we also aim to localize the object that appears OOC in the image. Recall that we generate counterfactual images that are in context by removing the candidate OOC object(s) from the scene. If removing an object turns the image to be in context then we consider the object as the OOC object. The results are shown in table 5.

Qualitative Results. We present the qualitative results to show the OOC images, corresponding captions, and the response from the GPT-4. Qualitative results for the MIT-OOC and IJCAI22-COCO-OOC datasets are shown in figure 7 and figure 8, respectively. GPT-4 not only rates the commonness of a context but also provides a justification for the rating and considers alternative ratings in other situations. This justifies the non-binary notion of OOC-ness.

4.2. Analysis of the approach.

We analyze results in terms of discovering failure modes and characterizing various OOC scenarios. We consider the MIT-OOC for the analysis as it contains natural images with diverse OOC scenarios.

False positives analysis. A reliable OOC detector is expected not to confuse in-context images as OOC. To evaluate this property, we estimate the false positive rate, i.e., how likely an in-context image is confused as OOC, of our approach while presented with in-context images. We consider the COCO 2014 validation set [28] as the set of in-context images. We achieve a false positive rate of 17.04%. Though these images are in context, some scenarios can be less common and thus deemed as OOC by the FM. Note that we achieve a low false positive rate without learning COCO-specific context relations. Complementary to false negatives, we achieve 9.18% and 12.74% false negative rates on MIT-OOC and COCO-OOC datasets, respectively.

Dataset	Accuracy (%)
MIT-OOC	63.32
IJCAI22-COCO-OOC	83.72

Table 5. Performance for localizing OOC objects in OOC images using our pipeline.

Error analysis. Our approach for detecting OOC images has two main steps - image captioning using BLIP-2, and querying the GPT-4 to evaluate whether the caption describes a usual context. Both of these steps can fail independently. We present a detailed error analysis for the MIT-OOC dataset which has natural images. The failures can be partitioned into three categories based on the source of the error. In the first category (Figure 4), the generated caption ignores the OOC objects and focuses on some other aspect of the scene. A more complete caption generation model could be used to deal with these kinds of failures. The second category of failures (Figure 5) arises due to the caption missing the surprising attribute such as the size being too large or too small, or the count of some objects being too high. The third category of failures (Figure 6) are those where the captions were generated correctly and informatively, but the GPT-4 foundation model did not find the description to be very surprising due to the nuanced nature of OOC. Manual inspection of these images revealed that the contexts in these images can be argued to be only somewhat surprising despite their inclusion in the MIT-OOC dataset.

Limitations and future directions. As we rely on FMs for image captioning to capture the object attributes and relations between objects, and on large language models for determining whether the caption describes a common context, our approach is limited by the accuracy of these models. Our failure analysis also illustrates these cases. Our pipeline will be able to leverage further advances in these FMs directly for improved zero-shot OOC detection.

5. Conclusion

We present a novel zero-shot approach for OOC detection using foundation models. Our method does not rely on learning dataset-specific context relations or manual definitions of context, but instead leverages the vast world knowledge encoded in large language models, such as GPT-4. Our experiments on the MIT-OOC and IJCAI22-COCO-OOC datasets show the efficacy of our approach. Notably, on the MIT-OOC, the results show significant improvements in detecting OOC images compared to the baseline VQA model

	No Error	OOC object missed in caption	OOC object attribute missed in caption	Caption is correct but nuanced OOC
Percentage	90.82	3.28	1.97	3.93

Table 6. Distribution of errors across different categories for the MIT-OOC dataset samples.

Figure 4. Samples from the MIT-OOC dataset along with the BLIP-2 generation caption. We fail to detect OOCs in these cases because the caption ignores the OOC aspect of the image (crashing of the car into brick wall in 087, snake on top shelf in 063, the large cup-shaped porches on the side of the building in 134, and the sofa in bus stand in 219.)

Figure 5. Samples from the MIT-OOC dataset along with the BLIP-2 captions for cases that failed due to the quantitative nature of out-of-context. The caption for 104 identifies the large sign on the chair but does not identify that the chair is quite large (comparable to a nearby tree and much larger than the cannon to the right of it). The caption for 068 identifies many toilet rolls but not a surprisingly large number of those. The caption for 189 identifies that the motorcycle is small but it does not register that it is surprisingly small.

Figure 6. Samples from the MIT-OOC dataset along with the BLIP-2 generated caption for cases that failed due to the nuanced nature of OOC despite the caption including the OOC object. Planes flying near the beach or sofa being on the side of the street is not very unusual in some places. The bench in 135 has chairs facing on different sides and this detail is not captured by the caption. The chair in 050 is more like a sofa and hence, unusual on a beach, and the caption describes it as a chair.

Caption: a cow crossing a street in a city with cars and buildings

GPT-4 response: I would rate it as a 2/10. It's quite an uncommon sight in most cities, but there may be certain urban areas in the world where it's not entirely unheard of.

Caption: a car is stuck in a flooded tunnel with a man standing next to it

GPT-4 response: This is difficult to quantify on a scale of 0 to 10 as it depends on factors such as geographical location, climate, and infrastructure. However, in general, this situation would not be very common, so it might be around a 2 or 3 out of 10.

Caption: a small car is driving on the water with people in it

GPT-4 response: I would say this is a 1 on the scale, as it is very uncommon for a small car to drive on the water. However, there are some specially designed amphibious vehicles that can drive on water, but they are rare and not typically used by the general public.

Figure 7. Qualitative results on MIT-OOC [11] showing the BLIP-2 captions and GPT-4 responses.

Caption: a laptop computer sitting on the side of a street

GPT-4 response: On a scale between 0 to 10, I would say this is a 2 because it is quite uncommon to see a laptop computer sitting on the side of a street.

Caption: two men riding horses on the beach with a suitcase in the background

GPT-4 response: I would say this is quite uncommon, so I would give it a 2 on a scale between 0 to 10.

Caption: a group of children playing soccer on a field with a pizza

GPT-4 response: I would say this is around a 4 on the scale. While children playing soccer on a field is common, having a pizza there at the same time is less likely.

Figure 8. Qualitative results on IJCAI22-COCO-OOC [1] showing the BLIP-2 captions and GPT-4 responses.

and state-of-the-art methods that rely on learning contextual relations. Additionally, our method often successfully localizes OOC objects in images. We perform extensive error analysis to determine the limitations of our approach and gain insights into areas where improvements can be made. Failures are categorized into three types: 1) the OOC object is missed in the image caption, 2) surprising attributes of OOC objects are missed in the caption, and 3) the GPT-4 model failed to find the scene surprising due to the nuanced nature of OOC. Future work may include more accurate captioning models and more nuanced prompts to GPT-4.

Acknowledgments

This work was supported in part by the U.S. Air Force and DARPA under Contract No. FA8750-23-C-0519, the U.S. Army Research Laboratory Cooperative Research Agreement W911NF-17-2-0196, the U.S. Army Combat Capabilities Development Command Army Research Laboratory under Support Agreement No. USMA 21050, DARPA under Support Agreement No. USMA 23004, U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research program, Reaching a New Energy Sciences Workforce under Award Number DE-SC0024576, DARPA under awards HR00112420004 and HR00112490420. The opinions, findings, and conclusions expressed in this paper are those of the authors and do not reflect the position of the United States Department of Defense or the United States Government.

References

[1] Manoj Acharya, Anirban Roy, Kaushik Koneripalli, Susmit Jha, Christopher Kanan, and Ajay Divakaran. Detecting out-of-context objects using graph context reasoning network. In *IJCAI*, 2022. 1, 2, 3, 5, 6, 8

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeuRIPS*, 2022. 3, 4

[3] Ehud Barnea and Ohad Ben-Shahar. Exploring the bounds of the utility of context for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7412–7420, 2019. 1

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 1, 3

[5] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, 2016. 1

[6] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, pages 1–17, 2021. 5

[7] Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, and Gabriel Kreiman. When pigs fly: Contextual reasoning in synthetic and natural scenes. *arXiv preprint arXiv:2104.02215*, 2021. 1, 2, 3

[8] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022. 3

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020. 3

[10] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. A tree-based context model for object recognition. *IEEE transactions on pattern analysis and machine intelligence*, 34(2):240–252, 2011. 5

[11] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012. 1, 2, 3, 5, 6, 8

[12] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 3

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng, Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose visionlanguage models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 5

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 3

[15] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2601–2610, 2019. 3

[16] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5

[17] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018. 1

[18] Amine Elhafsi, Rohan Sinha, Christopher Agia, Edward Schmerling, Issa AD Nesnas, and Marco Pavone. Semantic anomaly detection with large language models. *Autonomous Robots*, pages 1–21, 2023. 4

[19] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven CH Hoi. From images to textual prompts: Zero-shot vqa with frozen large language models. *arXiv preprint arXiv:2212.10846*, 2022. 3, 4

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5

[22] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022. 4

[23] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 1, 3

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 4

[25] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 3

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 3, 4, 5

[27] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 3

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 4, 5, 6

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 5

[30] David Madras and Richard Zemel. Identifying and benchmarking natural out-of-context prediction problems. In *NeuRIPS*, 2021. 1, 3

[31] Oscar Mañas, Pau Rodriguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. *arXiv preprint arXiv:2210.07179*, 2022. 3

[32] Gary Marcus and Ernest Davis. Experiments testing gpt-3's ability at commonsense reasoning: results, 2020. 4

[33] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007. 1, 3

[34] OpenAI. Gpt-4 system card, 2023. 1, 3, 4

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[36] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 1, 4, 5

[38] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018. 1, 3

[39] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 4

[40] Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Tianyi Zhou, and Daxin Jiang. Large language models are strong zero-shot retriever. *arXiv preprint arXiv:2304.14233*, 2023. 4

[41] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022. 4

[42] Jin Sun and David W Jacobs. Seeing what is not there: Learning context to determine where objects are missing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5716–5724, 2017. 1, 3

[43] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*, 2022. 3, 4

[44] Antonio Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2):169–191, 2003. 1

[45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[46] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*, 2022. 4

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[48] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5

[49] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12985–12994, 2020. 1, 3, 5

[50] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*, 1(6):8, 2021. 3

[51] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3

[52] Yimeng Zhang and Tsuhan Chen. Efficient inference for fully-connected crfs with stationarity. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 582–589. IEEE, 2012. 1, 3